Fast-rate and optimistic-rate error bounds for ℓ_1-regularized regression

Rina Foygel and Nathan Srebro

August 2, 2011

Abstract

We consider the prediction error of linear regression with ℓ_1 regularization when the number of covariates p is large relative to the sample size n. When the model is k-sparse and well-specified, and restricted isometry or similar conditions hold, the excess squared-error in prediction can be bounded on the order of $\frac{\sigma^2 k \log(p)}{n}$, where σ^2 is the noise variance. Although these conditions are close to necessary for accurate *recovery* of the true coefficient vector, it is possible to guarantee good predictive accuracy under much milder conditions, avoiding the restricted isometry condition, but only ensuring an excess error bound of order $\frac{k \log(p)}{n} + \sigma\sqrt{\frac{k \log(p)}{n}}$. Here we show that this is indeed the best bound possible (up to logarithmic factors) without introducing stronger assumptions similar to restricted isometry.

1 Introduction

We consider a random design linear regression problem with p covariates:

$$y = x^T \beta^* + z$$

where $x \in \mathbb{R}^p$ are random covariates with covariance matrix Σ, z is random noise with $\mathbb{E}\left[z^2\right] = \sigma^2$, and $\beta^* \in \mathbb{R}^p$ are the regression coefficients. For simplicity we take the response to be normalized, $\mathbb{E}\left[y^2\right] = 1$ (otherwise all results scale accordingly).

We consider the problem of minimizing the prediction error

$$\mathbb{E}\left[\left(y - x^T\beta\right)^2\right]$$

based on an i.i.d. sample $\left(x^{(1)}, y^{(1)}\right) \ldots, \left(x^{(n)}, y^{(n)}\right)$ using ℓ_1-regularized regression:

$$\hat{\beta}^B \doteq \arg\min_{\|\beta\|_1 \le B} \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2 .$$

Note that up to some unknown and data-dependent correspondence between B and λ, this is the same as

$$\hat{\beta}_\lambda \doteq \arg\min_\beta \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2 + \lambda \|\beta\|_1 ,$$

also known as Lasso regression [Tibshirani, 1996].

Suppose that the covariates are 1-bounded, and that $\max_i \left|y^{(i)}\right| \leq \mathbf{O}\left(\log(np)\right)$ (for instance, this is true with high probability in the Gaussian setting). Then, by Srebro et al. [2010], with high probability over the sample, for any fixed β^* with $\|\beta^*\|_1 \leq B$, excess squared-error under ℓ_1-regularized regression is bounded as

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] - \mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right] = \mathbf{O}\left(\frac{(1+B)^2\log(p)}{n/\log^3(n)} + \sqrt{\frac{(1+B)^2\log(p)}{n/\log^3(n)} \cdot \mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right]}\right) . \tag{1}$$

This result does not require any conditions on the correlation between the covariates, or on the nature of the "noise" $y - x^T\beta^*$, aside from the mild bound on $\max_i \left|y^{(i)}\right|$. In particular, this noise is not required to be independent from x. We believe also that this result would hold for subgaussian x's (rather than our current stronger assumption that the x's are 1-bounded).

We can apply this result to the sparse regression setting, with some mild additional assumptions. Suppose that we are interested in comparing to a sparse predictor on an unknown support $J^* \subset [p]$, with $|J^*| \leq k$. We now place a lower-bound eigenvalue assumption on this support J^* only:

$$\lambda_{\min}\left(\mathbb{E}\left[x_{J^*}x_{J^*}^T\right]\right) \geq \lambda_1 > 0 , \tag{2}$$

where $x_{J^*} = (x_j : j \in J^*)$ is the random vector consisting of those covariates x_j for which β_j^* is nonzero. This assumption is strictly weaker than the restricted isometry property (RIP) conditions in the compressed sensing literature, which require an upper-bound assumption as well, and require the eigenvalue bounds to hold for all sets $J \subset [p]$ of bounded size, in addition to the true support J^*.

We fix the scale of the problem by assuming $\mathbb{E}\left[y^2\right] = 1$. Now consider a predictor β^* with support in S^*, which is better than the zero predictor — that is, $\mathbb{E}\left[(y - x^T\beta^*)^2\right] \leq \mathbb{E}\left[(y - x^T\mathbf{0}_p)^2\right] = \mathbb{E}\left[y^2\right]$=1. We now show that $\|\beta^*\|_1 = \mathbf{O}\left(\sqrt{k\lambda_1^{-1}}\right)$. We first bound $\|\beta^*\|_2^2$, by observing that

$$\begin{aligned}
&\|\beta^*\|_2^2 \cdot \lambda_1 \leq (\beta^*)^T\mathbb{E}\left[xx^T\right]\beta^* = \mathbb{E}\left[\left(x^T\beta^*\right)^2\right] = \mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right] - 2\mathbb{E}\left[y \cdot \left(y - x^T\beta^*\right)\right] + \mathbb{E}\left[y^2\right] \\
&\leq 2\mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right] + 2\mathbb{E}\left[y^2\right] \leq 4 .
\end{aligned}$$

We then have

$$\|\beta^*\|_1 \leq \sqrt{k}\|\beta^*\|_2 \leq \sqrt{k \cdot 4\lambda_1^{-1}} = \mathbf{O}\left(\sqrt{k\lambda_1^{-1}}\right) .$$

Therefore, with high probability,

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] - \mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right] = \mathbf{O}\left(\frac{k\log(p)}{\lambda_1 n/\log^3(n)} + \sqrt{\frac{k\log(p)}{\lambda_1 n/\log^3(n)} \cdot \mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right]}\right) , \tag{3}$$

under the assumption that the x's are 1-bounded and $\max_i \left|y^{(i)}\right| \leq \mathbf{O}(np)$. Therefore, to guarantee a bound of ϵ on the excess prediction error, the required sample complexity is

$$n = \mathbf{\Theta}\left(\frac{k\log(p)}{\lambda_1\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot \log^3\left(k/\lambda_1\epsilon\right)\right) , \tag{4}$$

where $\sigma^2 = \mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right]$ is the magnitude of the noise. This sample complexity follows an "optimistic rate": in the noisy setting, if we would like to ensure a bound ϵ on excess error which is small relative to σ^2,

then the required sample complexity is then $n = \mathbf{\Theta}\left(\epsilon^{-2}\right)$, but on the the other hand, in the noiseless setting (i.e. when $y = x^T\beta^*$), or if the bound on excess error ϵ is not much smaller than σ^2, then we require only $n = \mathbf{\Theta}\left(\epsilon^{-1}\right)$. We emphasize that this result does not assume that the linear model is a true model or require independent noise.

In contrast, results on sparse vector recovery from the compressed sensing framework [Candes and Tao, 2005, Bickel et al., 2009, Koltchinskii, 2009, Cai et al., 2009] provide stronger guarantees in a similar setting, using either ℓ_1-regularized regression or the Dantzig selector, given by

$$\hat{\beta}^{DS}_{\lambda} = \arg\min \max_i \left| y^{(i)} - x^{(i)T}\beta \right| + \lambda\|\beta\|_1 \ .$$

These stronger results require several additional specialized assumptions, including the requirement that the noise must be independent from the signal. Existing results are stated either in the deterministic or random covariates setting, but can in general be translated to a random Gaussian setting. We restrict our attention to ℓ_1-regularized regression when the covariates are i.i.d. multivariate Gaussian with zero mean: $x^{(i)} \overset{i.i.d.}{\sim} N(0, \Sigma)$. We now summarize this setting (with some simplifications), and compare it to the optimistic-rate results discussed above.

- **Well-specified model with independent subgaussian noise:** Response $y^{(i)}$ is given by $y^{(i)} = x^{(i)T}\beta^* + \sigma z^{(i)}$ for a true predictor β^* satisfying $\|\beta^*\|_1 \leq B$, and $z^{(i)}$ is a subgaussian or subexponential noise term with unit variance, and is independent from $x^{(i)}$.
  The main additional requirement here is that noise z is independent of x. This in particular implies that β^* is the optimal regressor. Note that in order to obtain the optimistic-rate guarantee (3), no such assumption is necessary, and β^* can be a non-optimal regressor chosen for its sparsity or eigenvalue properties.

- **Sparsity:** β^* is k-sparse, meaning that it has (at most) k non-zero entries.
  To obtain the optimistic-rate guarantee as stated originally in (1), we can relax this requirement and only assume that β^* has low ℓ_1-norm.

- **Restricted eigenvalues:** There exists a $\kappa \doteq \kappa(k, 3) > 0$, such that for any $J \subset [p]$ with $|J| \leq k$, for any nonzero $\beta \in \mathbb{R}^p$ with $\|\beta_{\overline{J}}\|_1 \leq 3\|\beta_J\|_1$,

  $$\beta^T\Sigma\beta \geq \kappa\|\beta_J\|_2^2 \ . \tag{5}$$

  This restricted eigenvalue condition is implied by a stronger condition:
  Restricted isometry: Suppose that $\delta_{2k} + 3\theta_{k,2k} < 1$, where $v^T\Sigma v \in (1 \pm \delta_{2k})\|v\|_2^2$ for all $2k$-sparse vectors v, and $\left|v^T\Sigma w\right| \leq \theta_{k,2k}\|v\|_2\|w\|_2$ for all k-sparse v and $2k$-sparse w with disjoint supports. Then $\kappa \doteq \sqrt{1-\delta_{2k}}\left(1 - \frac{3\theta_{k,2k}}{1-\delta_{2k}}\right)$ satisfies the restricted eigenvalue condition above.
  To obtain the optimistic-rate guarantee (3) under the sparsity assumption, we required an eigenvalue condition (2) on $\Sigma_{\mathrm{Support}(\beta^*)}$ only, which is strictly weaker than the restricted eigenvalue and restricted isometry assumptions.

Under these assumptions, with κ defined as in (5), the following guarantees hold with high probability, by Theorem 7.2 of Bickel et al. [2009]:

$$\text{Sparse and accurate estimation of } \beta^*\text{: } \left\|\hat{\beta}^B - \beta^*\right\|_1 = \mathbf{O}\left(\frac{\sigma k}{\kappa^2} \cdot \sqrt{\frac{\log(p)}{n}}\right) \text{ , and } \|\hat{\beta}^B\|_0 = \mathbf{O}\left(k\right) \ .$$

$$\text{Bounded excess prediction error: } \mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] = \mathbb{E}\left[\left(y - x^T\beta^*\right)^2\right] + \mathbf{O}\left(\frac{\sigma^2 k\log(p)}{\kappa^2 n}\right) \ . \tag{6}$$

This corresponds to a sample complexity of

$$n = \mathbf{\Theta}\left(\frac{\sigma^2 k \log(p)}{\kappa^2 \epsilon}\right) , \tag{7}$$

to ensure an excess error bound of ϵ. It is crucial to note that the error bound (and the sample complexity) scales with the magnitude of the noise, σ^2, rather than to the (unit) magnitude of the signal. In particular, in a noiseless setting, the results above guarantee a zero-error reconstruction of β^*, in contrast to the "optimistic rate" result (3) where no such guarantee is given. Furthermore, in the noisy setting, the compressed sensing guarantees give a "fast rate" result, since the sample complexity scales with $\frac{1}{\epsilon}$ rather than with $\frac{1}{\epsilon^2}$.

In this compressed sensing framework, the guarantees on predictive error follow from a stronger guarantee on the accurate recovery of β^*, and in particular, the recovery of the true support of β^*. In order for this to be possible, it is of course necessary to be able to distinguish between pairs or small sets of covariates. In particular, some sort of restricted isometry assumption is clearly necessary for bounding error in recovering β^* (otherwise, the "best" β^* might not be unique). However, if the goal is merely low error in prediction — that is, we would like accuracy in calculating $x^T\beta^*$, rather than in recovering β^* — then perhaps this assumption could be weakened. For example, if a covariate is duplicated in the model, then it will not be possible to distinguish between the two when attempting to recover the true support; however, adding duplicated covariates to a model will have no effect on the problem of prediction.

More generally, we are interested in whether the properties that are necessary for the (unique) recovery of β^*, are also necessary to obtain strong bounds on excess prediction error, and in the role of the assumptions that separate the "optimistic rate", unit-scale error bounds of Srebro et al. [2010] from the "fast rate" error bounds in the compressed sensing literature, which scale with the magnitude of the noise. Below, we show that, if we remove either the sparsity assumption (while still assuming that β^* has low ℓ_1-norm) or the restricted isometry assumption from the compressed sensing framework described above, then up to logarithmic factors, the "optimistic rate" bound on excess prediction error, given in (3), is the best possible bound. In particular, this implies that, even in the noiseless setting, we cannot achieve zero error in prediction, without stronger assumptions.

2 Results

First, we ask whether we can relax the assumption of a sparse true coefficient vector to an assumption on its ℓ_1-norm, but still guarantee a fast-rate bound on excess error. Specifically, we consider the question of bounding excess prediction error, in the well-specified Gaussian setting where the restricted eigenvalue assumption holds, assuming only an ℓ_1-norm bound on the true vector of coefficients.

Our first result shows that, up to logarithmic factors, the optimistic-rate error bound (3) is the best possible rate under these conditions. For simplicity, we will consider the case of completely independent covariates, $x \sim N(0, \mathbf{I}_p)$. In particular, this ensures that the restricted eigenvalue assumption is satisfied. To place the problem on a unit scale (or rather, to bound the scale away from zero and away from infinity), we consider only true coefficient vectors β^* satisfying

$$\frac{1}{2} \leq \mathbb{E}\left[\left(x^T\beta^*\right)^2\right]^{1/2} = \|\beta^*\|_2 \leq \|\beta^*\|_1 \leq 1 .$$

Theorem 1. *Fix any $n \geq 30$, $p \geq 3n$, and $\sigma \geq 0$. Then there exists a $\beta^* \in \mathbb{R}^p$ with $\frac{1}{2} \leq \|\beta^*\|_2 \leq \|\beta^*\|_1 \leq 1$, such that for any sample, for all $B \geq 0$,*

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{1}{32n\log^2(3n)} .$$

Additionally, if $100 \leq \sqrt{n}/\sigma \leq p$*, then with probability at least* $\frac{1}{2}$ *over the sample, for all* $B \geq 0$*,*

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{\sigma}{102400\sqrt{n}\log^2\left(\max\left\{3n, \lceil\sqrt{n}/\sigma\rceil\right\}\right)} .$$

Here $\hat{\beta}^B = \arg\min_{\|\beta\|_1 \leq B} \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$*, where* $\left(x^{(i)}, y^{(i)}\right)$ *are i.i.d. samples from the multivariate Gaussian distribution defined by drawing* $x^{(i)} \sim N(0, \mathbf{I}_p)$ *and* $y^{(i)} \sim N\left(x^{(i)T}\beta^*, \sigma^2\right)$*. The expectations are taken over a new sample* (x, y) *drawn from the same distribution, independently of the training set* $\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$*. (For each* $B \geq 0$*, if* $\hat{\beta}^B$ *is not unique, then we show that the inequalities hold for some choice of* $\hat{\beta}^B$*.)*

Next, we ask whether the restricted eigenvalue (or restricted isometry) assumption is necessary for a fast-rate bound on excess error, in the well-specified Gaussian setting where the sparsity assumption holds.

Our second result shows that, up to logarithmic factors, the optimistic-rate error bound (3) is the best possible rate under these conditions. For simplicity, we restrict our attention to 2-sparse true coefficient vectors. We also only consider covariance matrices Σ such that $\Sigma_{J^*} = \mathbf{I}_{J^*}$, where $J^* = \mathrm{Support}\,(\beta^*)$. That is, ensuring the restricted isometry property on the true support only, is not sufficient for a fast-rate bound on excess error.

To avoid issues of scaling, we restrict our attention to covariance matrices Σ with $\|\Sigma\|_{sp} \leq 2$, and to true coefficient vectors β^* satisfying

$$\frac{1}{2} \leq \mathbb{E}\left[\left(x^T\beta^*\right)^2\right]^{1/2} = \sqrt{\beta^{*T}\Sigma\beta^*} = \|\beta^*\|_2 \leq \|\beta^*\|_1 \leq 1 ,$$

where we make use of the fact that $\Sigma_{J^*} = \mathbf{I}_{J^*}$ to obtain the second equality.

Theorem 2. *Fix any* $n \geq 30$*,* $p \geq 3n$*, and* $\sigma \geq 0$*. Then there exists a 2-sparse* $\beta^* \in \mathbb{R}^p$ *with* $\frac{1}{2} \leq \|\beta^*\|_2 \leq \|\beta^*\|_1 \leq 1$*, and a positive semi-definite* $\Sigma \in \mathbb{R}^{p\times p}$ *with* $\|\Sigma\|_{sp} \leq 2$ *and* $\Sigma_{\mathrm{Support}(\beta^*)} = \mathbf{I}_{\mathrm{Support}(\beta^*)}$*, such that for any sample, for all* $B \geq 0$*,*

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{1}{288n\log^2(3n)} .$$

Additionally, if $100 \leq \sqrt{n}/\sigma \leq p - 3$*, then with probability at least* $\frac{1}{2}$ *over the sample, for all* $B \geq 0$*,*

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{\sigma}{409600\sqrt{n}\log^2\left(\max\left\{3n, \lceil\sqrt{n}/\sigma\rceil\right\}\right)} .$$

Here $\hat{\beta}^B = \arg\min_{\|\beta\|_1 \leq B} \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$*, where* $\left(x^{(i)}, y^{(i)}\right)$ *are i.i.d. samples from the multivariate Gaussian distribution defined by drawing* $x^{(i)} \sim N(0, \Sigma)$ *and* $y^{(i)} \sim N\left(x^{(i)T}\beta^*, \sigma^2\right)$*. The expectations are taken over a new sample* (x, y) *drawn from the same distribution, independently of the training set* $\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$*. (For each* $B \geq 0$*, if* $\hat{\beta}^B$ *is not unique, then we show that the inequalities hold for some choice of* $\hat{\beta}^B$*.)*

In particular, Theorem 2 shows that without placing any assumptions on the covariates outside of $\mathrm{Support}\,(\beta^*)$, we cannot guarantee a bound on excess error that is better than the optimistic rate obtained by Srebro et al. [2010] from concentration bounds, up to logarithmic factors.

3 Proofs

We begin by defining a class of predictors that are optimal with respect to the squared-error loss and the ℓ_1-norm regularizer:

Definition 1. Given $y^{(i)} \in \mathbb{R}$ and $x^{(i)} \in \mathbb{R}^p$ for $i = 1, \ldots, n$, a predictor $\tilde{\beta} \in \mathbb{R}^p$ is Pareto-optimal (with respect to empirical squared-error and ℓ_1-norm) if it satisfies

$$\sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2 \leq \sum_i \left(y^{(i)} - x^{(i)T}\tilde{\beta}\right)^2 \;\Rightarrow\; \|\beta\|_1 \geq \|\tilde{\beta}\|_1 \;,$$

that is, if we cannot improve its empirical squared error without increasing its ℓ_1-norm, and vice versa.

The following lemma states a well-known property of ℓ_1-regularized regression; we include a proof for completeness.

Lemma 1. *For any $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}$ and $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$, for any $B \geq 0$, the class*

$$\mathcal{B}_B \doteq \arg\min_{\|\beta\|_1 \leq B} \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$$

must contain a predictor $\hat{\beta}^B$ that is Pareto-optimal and satisfies $\|\hat{\beta}^B\|_0 \leq n$.

Proof. Let $\mathrm{Err}_B = \inf_{\|\beta\|_1 \leq B} \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$. Since $\{\|\beta\|_1 \leq B\}$ is a compact set, this infimum is attained by some β with $\|\beta\|_1 \leq B$. Now define

$$B' = \inf\left\{\|\beta\|_1 : \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2 \leq \mathrm{Err}_B\right\} \leq B \;.$$

Again, by compactness, this infimum is attained by some $\tilde{\beta}$. We then see that $\tilde{\beta}$ is Pareto-optimal by its construction. Finally, by Theorem 3 of Rosset et al. [2004], there exists a $\hat{\beta}^B \in \mathbb{R}^p$ such that $\|\hat{\beta}^B\|_0 \leq n$, $\|\hat{\beta}^B\|_1 \leq \|\tilde{\beta}\|_1$, and $X\hat{\beta}^B = X\tilde{\beta}$. This is sufficient. □

Next we state two additional lemmas, proved in the next section.

Lemma 2. *Fix n and p with $n \geq 30$ and $p \geq 3n$. Let $x^{(i)} \overset{i.i.d.}{\sim} N(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{p\times p}$, and let $\beta^* \in \mathbb{R}^p$ be fixed. Then with probability at least $1 - 2e^{-n\log(p)}$, for all $J \subset [p]$ with $|J| = n$,*

$$\left\|X_J^T X(\tilde{\beta} - \beta^*)\right\|_2 \leq \|\Sigma\|_{sp} \cdot 16\sqrt{2} \cdot n\log(p) \cdot \sqrt{(\tilde{\beta} - \beta^*)^T\Sigma(\tilde{\beta} - \beta^*)} \text{ for all } \tilde{\beta} \in \mathbb{R}^p \text{ with } \tilde{\beta}_{\overline{J}} = 0 \;, \quad (8)$$

where the matrix X has entries $X_{ij} = x_j^{(i)}$, and X_J consists of the columns of X indexed by $j \in J$.

Lemma 3. *Let $x^{(i)} \overset{i.i.d.}{\sim} N(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{p\times p}$, and let $z \in \mathbb{R}^n$ be fixed, with $\|z\|_2^2 \geq 0.5n$. Assume $\sqrt{n}/\sigma \geq 100$. Then with probability at least $1 - e^{-0.015\sigma^{-1}\sqrt{n}}$, for all $J_1 \subset [\lceil\sqrt{n}/\sigma\rceil]$ with $|J_1| \geq \frac{\sqrt{n}}{2\sigma}$,*

$$\left\|\mathrm{Proj}^{\perp}_{1_{J_1}}(X_{J_1}^T z)\right\|_2^2 \geq \frac{\lambda^2_{\min}(\Sigma_{J_1})\, n^{3/2}}{200\sigma} \;, \quad (9)$$

where the matrix X has entries $X_{ij} = x_j^{(i)}$, and X_{J_1} consists of the columns of X indexed by $j \in J_1$.

We now prove the theorems.

3.1 Proof of Theorem 1

Theorem 1. Fix any $n \geq 30$, $p \geq 3n$, and $\sigma \geq 0$. Then there exists a $\beta^* \in \mathbb{R}^p$ with $\frac{1}{2} \leq \|\beta^*\|_2 \leq \|\beta^*\|_1 \leq 1$, such that for any sample, for all $B \geq 0$,

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{1}{32n\log^2(3n)}\,. \tag{10}$$

Additionally, if $100 \leq \sqrt{n}/\sigma \leq p$, then with probability at least $\frac{1}{2}$ over the sample, for all $B \geq 0$,

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{\sigma}{102400\sqrt{n}\log^2\left(\max\left\{3n, \lceil\sqrt{n}/\sigma\rceil\right\}\right)}\,. \tag{11}$$

Here $\hat{\beta}^B = \arg\min_{\|\beta\|_1 \leq B}\sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$, where $\left(x^{(i)}, y^{(i)}\right)$ are i.i.d. samples from the multivariate Gaussian distribution defined by drawing $x^{(i)} \sim N(0, \mathbf{I}_p)$ and $y^{(i)} \sim N\left(x^{(i)T}\beta^*, \sigma^2\right)$. The expectations are taken over a new sample (x, y) drawn from the same distribution, independently of the training set $\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$. (For each $B \geq 0$, if $\hat{\beta}^B$ is not unique, then we show that the inequalities hold for some choice of $\hat{\beta}^B$.)

Proof. Let β^* be

$$\beta^*_j = \frac{1}{j \cdot 4\log p},\ j = 1, \ldots, p-1;\ \beta^*_p = \frac{1}{2}\,.$$

Note that $\|\beta^*\|_1 \leq 1$ and $\|\beta^*\|_2^2 \geq \frac{1}{4}$, and so the resulting distribution satisfies the desired assumptions.t

By Lemma 1, for any $B \geq 0$, the set $\arg\min_{\|\beta\|_1 \leq B}\sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$ must include a Pareto-optimal vector $\hat{\beta}^B$ with $\|\hat{\beta}^B\|_0 \leq n$. Therefore, it is sufficient to show that bounds (10) and (11) hold for all Pareto-optimal vectors $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$. We now prove these two bounds separately.

Proof of (10). For any $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$, we have

$$\left\|\hat{\beta} - \beta^*\right\|_2^2 \geq \sum_{j=1}^{p}\left(\hat{\beta} - \frac{1}{j \cdot 4\log p}\right)^2 \geq \sum_{j:\hat{\beta}_j = 0}\left(\frac{1}{j \cdot 4\log p}\right)^2 \geq \sum_{j=n+1}^{p}\left(\frac{1}{j \cdot 4\log p}\right)^2$$
$$\geq \frac{1}{16\log^2(p)}\int_{x=n+1}^{p}\frac{1}{x^2}\,dx = \frac{1}{16\log^2(p)}\left(\frac{1}{n+1} - \frac{1}{p}\right) \geq \frac{1}{32n\log^2(p)}\,.$$

This proves the claim when $p = 3n$. However, the claim is immediately true for any larger value of p, since we may add in an arbitrary number of zero covariates (and assign zero coefficients to these covariates), without affecting the results.

Proof of (11). By Lemma 1 of Laurent and Massart [2000], with probability at least $1 - e^{-0.0625n} \geq 0.75$, $\|z\|_2^2 \sim \chi^2_n \geq 0.5n$. For the remainder of the proof, we treat $z \in \mathbb{R}^n$ as a fixed vector, and assume $\|z\|_2^2 \geq 0.5n$.

Assume that (8) holds for all $J \subset [p]$ with $|J| = n$, and (9) holds for all $J_1 \subset [\lceil\sqrt{n}/\sigma\rceil]$ with $|J_1| \geq \sqrt{n}/2\sigma$. (By Lemmas 2 and 3, this is true with probability at least $1 - 2e^{-n\log(p)} - e^{-0.015\sigma^{-1}\sqrt{n}} \geq 0.75$.) Now choose any Pareto-optimal $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$.

Suppose that $\|\hat{\beta}-\beta^*\|_2^2 < \frac{\sigma}{102400\sqrt{n}\log^2(p)}$. First, we show that

$$\left|\left\{j \in [\lceil \sqrt{n}/\sigma \rceil] : \hat{\beta}_j > 0\right\}\right| \geq \frac{\sqrt{n}}{2\sigma} .$$

Suppose not. Then

$$\|\hat{\beta}-\beta^*\|_2^2 \geq \sum_{j\in[\lceil \sqrt{n}/\sigma \rceil]:\hat{\beta}_j \leq 0} (\beta_j^*)^2 = \frac{1}{16\log^2(p)} \sum_{j\in[\lceil \sqrt{n}/\sigma \rceil]:\hat{\beta}_j \leq 0} \frac{1}{j^2} \geq \frac{1}{16\log^2(p)} \sum_{j=\lceil \sqrt{n}/2\sigma \rceil}^{\lceil \sqrt{n}/\sigma \rceil} \frac{1}{j^2}$$
$$\geq \frac{1}{16\log^2(p)} \int_{x=\lceil \sqrt{n}/2\sigma \rceil}^{2\lceil \sqrt{n}/2\sigma \rceil} \frac{1}{x^2} dx = \frac{1}{16\log^2(p)} \left(\frac{1}{\lceil \sqrt{n}/2\sigma \rceil} - \frac{1}{2\lceil \sqrt{n}/2\sigma \rceil}\right) = \frac{1}{16\log^2(p)\cdot 2\lceil \sqrt{n}/2\sigma \rceil} \geq \frac{\sigma}{32\sqrt{n}\log^2(p)} .$$

This is a contradiction.

Now define $J_1 = \left\{j \in [\lceil \sqrt{n}/\sigma \rceil] : \hat{\beta}_j > 0\right\}$, and fix any $J \supset \text{Support}(\hat{\beta})$ with $|J| = n$. Since $\hat{\beta}$ is Pareto-optimal with positive entries $\hat{\beta}_j$ for all $j \in J_1$, we have

$$\frac{\partial}{\partial(\beta_{J_1})} \|\hat{\beta}\|_1 = \mathbf{1}_{J_1} .$$

Therefore, by the theory of Lagrange multipliers, we must have $X_{J_1}^T y - X_{J_1}^T X\hat{\beta} = C \cdot \mathbf{1}_{J_1}$, for some $C \in \mathbb{R}$. We then have

$$X_{J_1}^T X(\hat{\beta}-\beta^*) = \sigma \cdot X_{J_1}^T z - C \cdot \mathbf{1}_{J_1} . \tag{12}$$

By (8), the norm of the left-hand side of (12) can be bounded from above as

$$\left\|X_{J_1}^T X(\hat{\beta}-\beta^*)\right\|_2 \leq \left\|X_{J_1}^T X(\hat{\beta}-\beta^*)\right\|_2 \leq \|\Sigma\|_{sp} \cdot 16\sqrt{2} \cdot n\log(p) \cdot \sqrt{(\hat{\beta}-\beta^*)^T\Sigma(\hat{\beta}-\beta^*)} .$$

By (9), the norm of the right-hand side of (12) can be bounded from below as

$$\left\|\sigma \cdot X_{J_1}^T z - C \cdot \mathbf{1}_{J_1}\right\|_2 \geq \sigma \cdot \left\|\text{Proj}_{\mathbf{1}_{J_1}^\perp} X_{J_1}^T z\right\|_2 \geq \sigma\sqrt{\frac{\lambda_{\min}^2(\Sigma_{J_1})\, n^{3/2}}{200\sigma}} .$$

Therefore, returning to (12), we have

$$\|\Sigma\|_{sp} \cdot 16\sqrt{2} \cdot n\log(p) \cdot \sqrt{(\hat{\beta}-\beta^*)^T\Sigma(\hat{\beta}-\beta^*)} \geq \left\|X_{J_1}^T X(\hat{\beta}-\beta^*)\right\|_2$$
$$= \left\|\sigma \cdot X_{J_1}^T z - C \cdot \mathbf{1}_{J_1}\right\|_2 \geq \sigma\sqrt{\frac{\lambda_{\min}^2(\Sigma_{J_1})\, n^{3/2}}{200\sigma}} .$$

Therefore,

$$(\hat{\beta}-\beta^*)^T\Sigma(\hat{\beta}-\beta^*) \geq \frac{\sigma \cdot \lambda_{\min}^2(\Sigma_{J_1})}{102400\|\Sigma\|_{sp}^2 \cdot \sqrt{n}\log^2(p)} = \frac{\sigma}{102400\sqrt{n}\log^2(p)} .$$

This proves the claim when $p = \max\{3n, \lceil \sqrt{n}/\sigma \rceil\}$. As in the proof of (10), this is sufficient to prove the claim for any larger value of p.

□

3.2 Proof of Theorem 2

Theorem 2. Fix any $n \geq 30$, $p \geq 3n$, and $\sigma \geq 0$. Then there exists a 2-sparse $\beta^* \in \mathbb{R}^p$ with $\frac{1}{2} \leq \|\beta^*\|_2 \leq \|\beta^*\|_1 \leq 1$, and a positive semi-definite $\Sigma \in \mathbb{R}^{p \times p}$ with $\|\Sigma\|_{sp} \leq 2$ and $\Sigma_{\text{Support}(\beta^*)} = \mathbf{I}_{\text{Support}(\beta^*)}$, such that for any sample, for all $B \geq 0$,

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{1}{288n\log^2(3n)} \,. \tag{13}$$

Additionally, if $100 \leq \sqrt{n}/\sigma \leq p-3$, then with probability at least $\frac{1}{2}$ over the sample, for all $B \geq 0$,

$$\mathbb{E}\left[\left(y - x^T\hat{\beta}^B\right)^2\right] \geq \sigma^2 + \frac{\sigma}{409600\sqrt{n}\log^2\left(\max\left\{3n, \lceil\sqrt{n}/\sigma\rceil\right\}\right)} \,. \tag{14}$$

Here $\hat{\beta}^B = \arg\min_{\|\beta\|_1 \leq B} \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$, where $\left(x^{(i)}, y^{(i)}\right)$ are i.i.d. samples from the multivariate Gaussian distribution defined by drawing $x^{(i)} \sim N(0, \Sigma)$ and $y^{(i)} \sim N\left(x^{(i)T}\beta^*, \sigma^2\right)$. The expectations are taken over a new sample (x, y) drawn from the same distribution, independently of the training set $\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$. (For each $B \geq 0$, if $\hat{\beta}^B$ is not unique, then we show that the inequalities hold for some choice of $\hat{\beta}^B$.)

Proof. Let $w_1, w_2, u_1, \ldots, u_{p-3} \overset{iid}{\sim} N(0,1)$. Define

$$\tau = \frac{1}{4\log p} \cdot \left(\frac{1}{1}, \frac{1}{2}, \ldots, \frac{1}{p-3}\right) \in \mathbb{R}^{p-3} \,.$$

Since $p \geq 90$, $\|\tau\|_1 \leq \frac{1}{3}$ and $\|\tau\|_2^2 < \frac{1}{9\log^2(p)} \leq 0.01$. Now we define an additional covariate as a linear combination of the others:

$$v = \frac{1}{\sqrt{2}}(w_1 + w_2) \cdot \sqrt{1 - \|\tau\|_2^2} - u^T\tau \,.$$

Now define $x = (u_1, \ldots, u_{p-3}, v, w_1, w_2)$. Let $\Sigma = Cov(x)$, and note that $\sigma_{\max} = \|\Sigma\|_2 \leq 2$.

Define

$$\beta^*_{sparse} = \left(\mathbf{0}_{p-3}, 0, \frac{1}{2}, \frac{1}{2}\right), \; \beta^*_{dense} = \left(\frac{1}{\sqrt{2(1-\|\tau\|_2^2)}} \cdot \tau, \frac{1}{\sqrt{2(1-\|\tau\|_2^2)}}, 0, 0\right) \,.$$

and

$$y^{(i)} = \frac{1}{2}(w_1 + w_2) = x^{(i)T}\beta^*_{sparse} = x^{(i)T}\beta^*_{dense} \,.$$

Note that β^*_{sparse} and β^*_{dense} are both optimal predictors. Since $\|\beta^*_{sparse}\|_1 = 1$, ${\beta^*_{sparse}}^T\Sigma\beta^*_{sparse} = \|\beta^*_{sparse}\|_2^2 = \frac{1}{2}$, and β^*_{sparse} is 2-sparse, this distribution satisfies the desired assumptions. However,

$$\|\beta^*_{dense}\|_1 = \frac{1}{\sqrt{2(1-\|\tau\|_2^2)}}(1 + \|\tau\|_1) \approx \frac{4}{3\sqrt{2}} < 1 \,,$$

and so in a sense β^*_{dense} will be preferred to β^*_{sparse} in ℓ_1-regularized regression, thus leading to the same arguments as in the proof of Theorem 1.

By Lemma 1, for any $B \geq 0$, the set $\arg\min_{\|\beta\|_1 \leq B} \sum_i \left(y^{(i)} - x^{(i)T}\beta\right)^2$ must include a Pareto-optimal vector $\hat{\beta}^B$ with $\|\hat{\beta}^B\|_0 \leq n$. Therefore, it is sufficient to show that bounds (13) and (14) hold for all Pareto-optimal vectors $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$. For each such $\hat{\beta}$, we use the notation

$$\hat{\beta} = (\hat{\beta}_u, \hat{\beta}_{w_1}, \hat{\beta}_{w_2}, \hat{\beta}_v) \in \mathbb{R}^{p-3} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \, .$$

Observe that, by definition of the covariates,

$$\begin{aligned} &(\hat{\beta} - \beta^*_{sparse})^T \Sigma (\hat{\beta} - \beta^*_{sparse}) \\ &= \underbrace{\left\| \hat{\beta}_u - \tau \hat{\beta}_v \right\|_2^2}_{\text{(Term 1)}} + \underbrace{\left(\hat{\beta}_{w_1} + \frac{1}{\sqrt{2}} \sqrt{1 - \|\tau\|_2^2} \hat{\beta}_v - \frac{1}{\sqrt{2}} \right)^2}_{\text{(Term 2)}} + \underbrace{\left(\hat{\beta}_{w_2} + \frac{1}{\sqrt{2}} \sqrt{1 - \|\tau\|_2^2} \hat{\beta}_v - \frac{1}{\sqrt{2}} \right)^2}_{\text{(Term 3)}} \, . \end{aligned} \tag{15}$$

The remainder of the proof is organized as follows. First, we prove bounds (13) and (14) for any Pareto-optimal $\hat{\beta}$ with $\hat{\beta}_v \leq \frac{1}{3}$. Next, we prove the bound (13) for any Pareto-optimal $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$ and $\hat{\beta}_v > \frac{1}{3}$. Finally, we prove the bound (14) for any Pareto-optimal $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$ and $\hat{\beta}_v > \frac{1}{3}$.

Proof of (13) and (14) when $\hat{\beta}_v \leq \frac{1}{3}$. Consider any Pareto-optimal $\hat{\beta}$ with $\hat{\beta}_v \leq \frac{1}{3}$. First, suppose that $\hat{\beta}_{w_1}, \hat{\beta}_{w_2} \geq \frac{1}{2\sqrt{2}}$. Let

$$\tilde{\beta} = \left(\hat{\beta}_u + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \cdot \tau, \hat{\beta}_{w_1} - \tfrac{1}{2\sqrt{2}}, \hat{\beta}_{w_2} - \tfrac{1}{2\sqrt{2}}, \hat{\beta}_v + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \right) \, .$$

By the definition of the covariates, $x^{(i)T}\hat{\beta} = x^{(i)T}\tilde{\beta}$ for all i. We will now show that $\|\tilde{\beta}\|_1 < \|\hat{\beta}\|_1$. We have

$$\begin{aligned} \|\tilde{\beta}\|_1 &= \|\tilde{\beta}_u\|_1 + |\tilde{\beta}_{w_1}| + |\tilde{\beta}_{w_2}| + |\tilde{\beta}_v| \\ &= \left\| \hat{\beta}_u + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \cdot \tau \right\|_1 + \left| \hat{\beta}_{w_1} - \tfrac{1}{2\sqrt{2}} \right| + \left| \hat{\beta}_{w_2} - \tfrac{1}{2\sqrt{2}} \right| + \left| \hat{\beta}_v + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \right| \\ &= \left\| \hat{\beta}_u + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \cdot \tau \right\|_1 + \left| \hat{\beta}_{w_1} \right| + \left| \hat{\beta}_{w_2} \right| - \tfrac{1}{\sqrt{2}} + \left| \hat{\beta}_v + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \right| \\ &\leq \left\| \hat{\beta}_u \right\|_1 + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \cdot \|\tau\|_1 + \left| \hat{\beta}_{w_1} \right| + \left| \hat{\beta}_{w_2} \right| - \tfrac{1}{\sqrt{2}} + \left| \hat{\beta}_v \right| + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} \\ &= \|\hat{\beta}\|_1 - \tfrac{1}{\sqrt{2}} + \tfrac{1}{2\sqrt{1-\|\tau\|_2^2}} + \tfrac{\|\tau\|_1}{2\sqrt{1-\|\tau\|_2^2}} \leq \|\hat{\beta}\|_1 - \tfrac{1}{\sqrt{2}} + \tfrac{1}{2\sqrt{1-0.01^2}} + \tfrac{0.3}{2\sqrt{1-0.01^2}} \leq \|\hat{\beta}\|_1 - 0.05 \, . \end{aligned}$$

Therefore, this case leads to a contradiction, since we have constructed a coefficient vector $\tilde{\beta}$ with zero error on the training set, and lower ℓ_1-norm than $\hat{\beta}$. Therefore, we must have either $\hat{\beta}_{w_1} < \frac{1}{2\sqrt{2}}$ or $\hat{\beta}_{w_2} < \frac{1}{2\sqrt{2}}$. Without loss of generality, we assume $\hat{\beta}_{w_1} < \frac{1}{2\sqrt{2}}$.

Then

$$\hat{\beta}_{w_1} + \frac{1}{\sqrt{2}} \sqrt{1 - \|\tau\|_2^2} \hat{\beta}_v - \frac{1}{\sqrt{2}} < \frac{1}{2\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{1}{3} - \frac{1}{\sqrt{2}} \leq -\frac{1}{6\sqrt{2}} \, ,$$

and so by (Term 2) in (15) above,

$$(\hat{\beta} - \beta^*_{sparse})^T \Sigma (\hat{\beta} - \beta^*_{sparse}) \geq \left(\hat{\beta}_{w_1} + \frac{1}{\sqrt{2}} \sqrt{1 - \|\tau\|_2^2} \hat{\beta}_v - \frac{1}{\sqrt{2}} \right)^2 \geq \frac{1}{72} \, .$$

This is sufficient to show that both (13) and (14) are satisfied.

Proof of (13) when $\hat{\beta}_v > \frac{1}{3}$. Consider any Pareto-optimal $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$ and $\hat{\beta}_v > \frac{1}{3}$. We then have

$$\left\|\hat{\beta}_u - \tau\hat{\beta}_v\right\|_2^2 = \sum_{j=1}^{p-3}\left(\hat{\beta}_{u_j} - \frac{1}{4\log(p)}\cdot\frac{1}{j}\cdot\hat{\beta}_v\right)^2 \geq \sum_{j\in\{1,\ldots,p-3\}:\hat{\beta}_{u_j}=0}\left(\frac{1}{4\log(p)}\cdot\frac{1}{j}\cdot\hat{\beta}_v\right)^2$$

$$\geq \frac{\hat{\beta}_v^2}{16\log^2(p)}\cdot\sum_{j=n}^{p-3}\frac{1}{j^2} \geq \frac{1}{144\log^2(p)}\cdot\int_{x=n}^{p-3}\frac{1}{x^2}\,dx = \frac{1}{144\log^2(p)}\cdot\left(\frac{1}{n}-\frac{1}{p-3}\right) \geq \frac{1}{288n\log^2(p)}\,.$$

But, considering (Term 1) in (15) above, this proves that

$$(\hat{\beta}-\beta^*_{sparse})^T\Sigma(\hat{\beta}-\beta^*_{sparse}) \geq \frac{1}{288n\log^2(p)}\,.$$

This proves the claim when $p = 3n$. As in the proof of Theorem 1, this is sufficient to prove the claim for any larger value of p.

Proof of (14) when $\hat{\beta}_v > \frac{1}{3}$. By Lemma 1 of Laurent and Massart [2000], with probability at least $1 - e^{-0.0625n} \geq 0.75$, $\|z\|_2^2 \sim \chi^2_n \geq 0.5n$. For the remainder of the proof, we treat $z \in \mathbb{R}^n$ as a fixed vector, and assume $\|z\|_2^2 \geq 0.5n$.

Assume that (8) holds for all $J \subset [p]$ with $|J| = n$, and (9) holds for all $J_1 \subset [\lceil\sqrt{n}/\sigma\rceil]$ with $|J_1| \geq \sqrt{n}/2\sigma$. (By Lemmas 2 and 3, this is true with probability at least $1 - 2e^{-n\log(p)} - e^{-0.015\sigma^{-1}\sqrt{n}} \geq 0.75$.) Consider any Pareto-optimal $\hat{\beta}$ with $\|\hat{\beta}\|_0 \leq n$ and $\hat{\beta}_v > \frac{1}{3}$. First, suppose that

$$\left|\left\{j \in [\lceil\sqrt{n}/\sigma\rceil] : \hat{\beta}_j > 0\right\}\right| < \tfrac{\sqrt{n}}{2\sigma}\,.$$

Then

$$\|\hat{\beta}_u - \tau\hat{\beta}_v\|_2^2 = \sum_{j=1}^{p-3}\left(\hat{\beta}_{u_j} - \frac{1}{4\log(p)}\cdot\frac{1}{j}\cdot\hat{\beta}_v\right)^2 \geq \sum_{j\in[\lceil\sqrt{n}/\sigma\rceil]:\hat{\beta}_{u_j}\leq 0}\left(\frac{1}{4\log(p)}\cdot\frac{1}{j}\cdot\hat{\beta}_v\right)^2$$

$$= \tfrac{{\hat{\beta}_v}^2}{16\log^2(p)}\sum_{j\in[\lceil\sqrt{n}/\sigma\rceil]:\hat{\beta}_{u_j}\leq 0}\tfrac{1}{j^2} \geq \tfrac{1}{144\log^2(p)}\sum_{j=\lceil\sqrt{n}/2\sigma\rceil}^{\lceil\sqrt{n}/\sigma\rceil}\tfrac{1}{j^2}$$

$$\geq \tfrac{1}{144\log^2(p)}\int_{x=\lceil\sqrt{n}/2\sigma\rceil}^{2\lceil\sqrt{n}/2\sigma\rceil}\tfrac{1}{x^2}dx = \tfrac{1}{144\log^2(p)}\left(\tfrac{1}{\lceil\sqrt{n}/2\sigma\rceil} - \tfrac{1}{2\lceil\sqrt{n}/2\sigma\rceil}\right) = \tfrac{1}{144\log^2(p)\cdot 2\lceil\sqrt{n}/2\sigma\rceil} \geq \tfrac{\sigma}{288\sqrt{n}\log^2(p)}\,.$$

Considering (Term 1) in (15), this proves that

$$(\hat{\beta}-\beta^*_{sparse})^T\Sigma(\hat{\beta}-\beta^*_{sparse}) \geq \left\|\hat{\beta}_u - \tau\hat{\beta}_v\right\|_2^2 \geq \tfrac{\sigma}{288\sqrt{n}\log^2(p)}\,.$$

Next, suppose instead that

$$\left|\left\{j \in [\lceil\sqrt{n}/\sigma\rceil] : \hat{\beta}_j > 0\right\}\right| \geq \tfrac{\sqrt{n}}{2\sigma}\,.$$

Define $J_1 = \left\{ j \in [\lceil \sqrt{n}/\sigma \rceil] : \hat{\beta}_j > 0 \right\}$, and fix any $J \supset \text{Support}(\hat{\beta})$ with $|J| = n$. Since $\hat{\beta}$ is Pareto-optimal with positive entries $\hat{\beta}_j$ for all $j \in J_1$, we have

$$\frac{\partial}{\partial (\beta_{J_1})} \|\hat{\beta}\|_1 = \mathbf{1}_{J_1} \ .$$

Therefore, by the theory of Lagrange multipliers, we must have $X_{J_1}^T y - X_{J_1}^T X \hat{\beta} = C \cdot \mathbf{1}_{J_1}$, for some $C \in \mathbb{R}$. We then have

$$X_{J_1}^T X(\hat{\beta} - \beta^*) = \sigma \cdot X_{J_1}^T z - C \cdot \mathbf{1}_{J_1} \ . \tag{16}$$

By (8), the norm of the left-hand side of (16) can be bounded from above as

$$\left\| X_{J_1}^T X(\hat{\beta} - \beta^*) \right\|_2 \leq \left\| X_{J_1}^T X(\hat{\beta} - \beta^*) \right\|_2 \leq \|\Sigma\|_{sp} \cdot 16\sqrt{2} \cdot n \log(p) \cdot \sqrt{(\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*)} \ .$$

By (9), the norm of the right-hand side of (16) can be bounded from below as

$$\left\| \sigma \cdot X_{J_1}^T z - C \cdot \mathbf{1}_{J_1} \right\|_2 \geq \sigma \cdot \left\| \text{Proj}_{\mathbf{1}_{J_1}^\perp} X_{J_1}^T z \right\|_2 \geq \sigma \sqrt{\frac{\lambda_{\min}^2 (\Sigma_{J_1}) \, n^{3/2}}{200\sigma}} \ .$$

Therefore, returning to (16), we have

$$\begin{aligned} \|\Sigma\|_{sp} \cdot 16\sqrt{2} \cdot n \log(p) \cdot \sqrt{(\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*)} &\geq \left\| X_{J_1}^T X(\hat{\beta} - \beta^*) \right\|_2 \\ &= \left\| \sigma \cdot X_{J_1}^T z - C \cdot \mathbf{1}_{J_1} \right\|_2 \geq \sigma \sqrt{\frac{\lambda_{\min}^2 (\Sigma_{J_1}) \, n^{3/2}}{200\sigma}} \ . \end{aligned}$$

Therefore,

$$(\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*) \geq \frac{\sigma \cdot \lambda_{\min}^2 (\Sigma_{J_1})}{102400 \|\Sigma\|_{sp}^2 \cdot \sqrt{n} \log^2(p)} = \frac{\sigma}{409600 \sqrt{n} \log^2(p)} \ .$$

This proves the claim when $p = \max\{3n, \lceil \sqrt{n}/\sigma \rceil\}$. As in the proof of Theorem 1, this is sufficient to prove the claim for any larger value of p.

□

4 Proofs for Lemmas

4.1 Proof of Lemma 2

Fix any $J \subset [p]$ with $|J| = n$. We will show that, with probability at least $1 - 2e^{-2n \log(p)}$,

$$\left\| X_J^T X(\tilde{\beta} - \beta^*) \right\|_2 \leq \|\Sigma\|_{sp} \cdot 16\sqrt{2} \cdot n \log(p) \cdot \sqrt{(\tilde{\beta} - \beta^*)^T \Sigma (\tilde{\beta} - \beta^*)} \text{ for all } \tilde{\beta} \in \mathbb{R}^p \text{ with } \tilde{\beta}_{\overline{J}} = 0 \ .$$

Since there are $\binom{p}{n} \leq p^n$ choices for the set J, this will be sufficient to prove the lemma.

Reorder the covariates to write $\Sigma = \begin{pmatrix} \Sigma_{JJ} & \Sigma_{J\overline{J}} \\ \Sigma_{\overline{J}J} & \Sigma_{\overline{J}\overline{J}} \end{pmatrix}$. Choose a Cholesky decomposition

$$\Sigma = \begin{pmatrix} U & V \\ \mathbf{0} & W \end{pmatrix}^T \begin{pmatrix} U & V \\ \mathbf{0} & W \end{pmatrix} .$$

Let $a^{(i)} \sim N(0, \mathbf{I}_p)$. Then $\begin{pmatrix} U & V \\ \mathbf{0} & W \end{pmatrix}^T a^{(i)} \sim N(0, \Sigma)$, and so

$$\begin{pmatrix} X_J & X_{\overline{J}} \end{pmatrix} \stackrel{\mathcal{D}}{=} \begin{pmatrix} A_J U & A_J V + A_{\overline{J}} W \end{pmatrix} ,$$

where the matrix A has entries $A_{ij} = a_j^{(i)}$, and A_J consists of the columns of A indexed by $j \in J$. We then have

$$\begin{aligned} &X_J^T X(\tilde{\beta} - \beta^*) \stackrel{\mathcal{D}}{=} U^T A_J^T \left(A_J U(\tilde{\beta} - \beta^*)_J + (A_J V + A_{\overline{J}} W)(\tilde{\beta} - \beta^*)_{\overline{J}} \right) \\ &= U^T A_J^T \left(A_J U(\tilde{\beta} - \beta^*)_J - (A_J V + A_{\overline{J}} W)\beta^*_{\overline{J}} \right) = U^T A_J^T \left(A_J \left(U(\tilde{\beta} - \beta^*)_J - V\beta^*_{\overline{J}} \right) - A_{\overline{J}} W \beta^*_{\overline{J}} \right) \end{aligned}$$

Below, we will show that

$$\|A_J\|_{sp} \leq \sqrt{16n\log(p)} \text{ with probability at least } 1 - e^{-2n\log(p)} , \tag{17}$$

$$\text{and } \left\|A_{\overline{J}} W \beta^*_{\overline{J}}\right\|_2 \leq \sqrt{16n\log(p)} \cdot \left\|W\beta^*_{\overline{J}}\right\|_2 \text{ with probability at least } 1 - e^{-2n\log(p)} . \tag{18}$$

Assuming that these bounds hold. Then for any $\tilde{\beta} \in \mathbb{R}^p$ with $\tilde{\beta}_{\overline{J}} = 0$, we have

$$\begin{aligned} &\left\| U^T A_J^T \left(A_J \left(U(\tilde{\beta} - \beta^*)_J - V\beta^*_{\overline{J}} \right) - A_{\overline{J}} W \beta^*_{\overline{J}} \right) \right\|_2 \\ &\leq \|U\|_{sp} \cdot \|A_J\|_{sp} \cdot \left(\|A_J\|_{sp} \cdot \left\| U(\tilde{\beta} - \beta^*)_J - V\beta^*_{\overline{J}} \right\|_2 + \left\|A_{\overline{J}} W \beta^*_{\overline{J}}\right\|_2 \right) \\ &\leq \|\Sigma\|_{sp} \cdot \sqrt{16n\log(p)} \cdot \left(\sqrt{16n\log(p)} \cdot \left\| U(\tilde{\beta} - \beta^*)_J - V\beta^*_{\overline{J}} \right\|_2 + \sqrt{16n\log(p)} \cdot \left\|W\beta^*_{\overline{J}}\right\|_2 \right) \\ &= \|\Sigma\|_{sp} \cdot 16n\log(p) \cdot \left(\left\| U(\tilde{\beta} - \beta^*)_J - V\beta^*_{\overline{J}} \right\|_2 + \left\|W\beta^*_{\overline{J}}\right\|_2 \right) \\ &\leq \|\Sigma\|_{sp} \cdot 16n\log(p) \cdot \sqrt{2} \cdot \left(\left\| U(\tilde{\beta} - \beta^*)_J - V\beta^*_{\overline{J}} \right\|_2^2 + \left\|W\beta^*_{\overline{J}}\right\|_2^2 \right)^{1/2} \\ &= \|\Sigma\|_{sp} \cdot 16\sqrt{2} \cdot n\log(p) \cdot \sqrt{(\tilde{\beta} - \beta^*)^T \Sigma (\tilde{\beta} - \beta^*)} . \end{aligned}$$

We conclude by proving (17) and (18). We first prove (17) using a construction from Keshavan et al. [2010]. First, define $\mathcal{U} = \left\{ u \in \left(\frac{1}{8\sqrt{n}} \mathbb{Z} \right)^n : \|u\|_2 \leq 1 \right\}$. By Remark 5.1 in Keshavan et al. [2010],

$$\|A_J\|_{sp} \leq \sqrt{2} \sup_{u,v \in \mathcal{U}} \left| u^T A_J v \right| .$$

For any $u, v \in \mathcal{U}$,

$$u^T A_J v = \sum_{i=1}^n \sum_{j \in J} u_i v_j A_{ij} \sim N(0, \|u\|_2^2 \|v\|_2^2) ,$$

therefore,

$$Pr\left(\left| u^T A_J v \right| \geq \sqrt{8n\log(p)} \right) \leq Pr\left(|N(0,1)| \geq \sqrt{8n\log(p)} \right) \leq e^{-4n\log(p)} .$$

Furthermore, $|\mathcal{U}| \le (2\lceil 8\sqrt{n}\rceil + 1)^n \le p^n$. So,

$$Pr\left(\|A_J\|_{sp} \le \sqrt{16n\log(p)}\right) \le Pr\left(\left|u^T A_J v\right| \le \sqrt{8\cdot n\log(p)} \text{ for all } u,v \in \mathcal{U}\right)$$
$$\ge 1 - p^{2n}e^{-4n\log(p)} \ge 1 - e^{-2n\log(p)} .$$

Next we prove (18). We have

$$\left\|A_{\overline{J}} W\beta^*_{\overline{J}}\right\|_2^2 = \sum_i \left(\sum_{j\in\overline{J}} a_j^{(i)} \left(W\beta^*_{\overline{J}}\right)_j\right)^2 \stackrel{\mathcal{D}}{=} \left\|W\beta^*_{\overline{J}}\right\|_2^2 \cdot \chi^2_n .$$

By Lemma 1 of Laurent and Massart [2000], $Pr\left(\chi^2_n \ge 16n\log(p)\right) \le e^{-2n\log(p)}$. This is sufficient.

4.2 Proof of Lemma 3

Choose any $J_2 \subset J_1$ with $|J_2| = \lceil \sqrt{n}/2\sigma\rceil$. Observe that $\left\|\text{Proj}^{\perp}_{\mathbf{1}_{J_1}}(X^T_{J_1} z)\right\|_2^2 \ge \left\|\text{Proj}^{\perp}_{\mathbf{1}_{J_2}}(X^T_{J_2} z)\right\|_2^2$, and so it is sufficient to only consider the sets J_2 of size $\lceil \sqrt{n}/2\sigma\rceil$.

Fix any $J_2 \subset [\lceil \sqrt{n}/\sigma\rceil]$ with $|J_2| = \lceil \sqrt{n}/2\sigma\rceil$. Let $P \in \mathbb{R}^{\lceil \sqrt{n}/2\sigma\rceil \times \lceil \sqrt{n}/2\sigma\rceil}$ be the orthogonal projection matrix corresponding to $Proj^{\perp}_{\mathbf{1}_{J_2}}(\cdot)$. Write $P\Sigma_{J_2}P = AA^T$ for $A \in \mathbb{R}^{\lceil \sqrt{n}/2\sigma\rceil \times (\lceil \sqrt{n}/2\sigma\rceil - 1)}$. Then $(X^T_{J_2} z) \sim N(0, \|z\|_2^2 \cdot \Sigma_{J_2})$ and so $\text{Proj}^{\perp}_{\mathbf{1}_{J_2}}(X^T_{J_2} z) \sim N(0, \|z\|_2^2 \cdot P\Sigma_{J_2}P)$, and therefore $\text{Proj}^{\perp}_{\mathbf{1}_{J_2}}(X^T_{J_2} z) \stackrel{\mathcal{D}}{=} \|z\|_2^2 \cdot Au$ for $u \sim N\left(0, \mathbf{I}_{\lceil \sqrt{n}/2\sigma\rceil - 1}\right)$. By examining the definition of A, we see that $u^T(A^TA)u \ge \|u\|_2^2 \cdot \lambda^2_{\min}\left(\Sigma_{[\lceil \sqrt{n}/\sigma\rceil]}\right)$, therefore,

$$\|\text{Proj}^{\perp}_{\mathbf{1}_{J_2}}(X^T_{J_2} z)\|_2^2 \stackrel{\mathcal{D}}{=} \|z\|_2^2 \cdot \|Au\|_2^2 \stackrel{\mathcal{D}}{\ge} 0.5n \cdot \lambda^2_{\min}\left(\Sigma_{[\lceil \sqrt{n}/\sigma\rceil]}\right) \cdot \chi^2_{\lceil \sqrt{n}/2\sigma\rceil - 1} .$$

Furthermore, the number of such sets J_2 is bounded by $2^{\lceil \sqrt{n}/\sigma\rceil}$. By the chi-square tail bounds from Foygel and Drton [2010], using the assumption that $\sqrt{n}/\sigma \ge 100$, we have

$$Pr\left(\chi^2_{\lceil \sqrt{n}/2\sigma\rceil - 1} \le \frac{\sqrt{n}}{100\sigma}\right) \le Pr\left(\chi^2_{\lceil \sqrt{n}/2\sigma\rceil - 1} \le 0.02 \cdot \tfrac{50}{49} \cdot (\lceil \sqrt{n}/2\sigma\rceil - 1)\right)$$
$$\le \exp\left\{\tfrac{1}{2}(\lceil \sqrt{n}/2\sigma\rceil - 2)\left(1 - 0.02\cdot\tfrac{50}{49} + \log\left(0.02\cdot\tfrac{50}{49}\right)\right)\right\} \le \exp\left\{\tfrac{1}{2}\left(\lceil \sqrt{n}/2\sigma\rceil \cdot \tfrac{48}{50}\right)\left(1 - 0.02\cdot\tfrac{50}{49} + \log\left(0.02\cdot\tfrac{50}{49}\right)\right)\right\}$$
$$\le e^{-0.7084\lceil \sqrt{n}/\sigma\rceil} .$$

Therefore,

$$Pr\left(\exists J_1 \subset \left[\lceil\sqrt{n}\rceil\right], |J_1| \ge \tfrac{\sqrt{n}}{2},\ \|\text{Proj}^{\perp}_{\mathbf{1}_{J_1}}(X^T_{J_1} z)\|_2^2 \le n\lambda^2_{\min}\left(\Sigma_{[\lceil\sqrt{n}\rceil]}\right)\cdot\tfrac{\sqrt{n}}{200\sigma}\right)$$
$$\le Pr\left(\exists J_2 \subset [\lceil\sqrt{n}\rceil], |J_2| = \tfrac{\lceil\sqrt{n}\rceil}{2},\ \|\text{Proj}^{\perp}_{\mathbf{1}_{J_2}}(X^T_{J_2} z)\|_2^2 \le n\lambda^2_{\min}\left(\Sigma_{[\lceil\sqrt{n}\rceil]}\right)\cdot\tfrac{\sqrt{n}}{200\sigma}\right)$$
$$\le 2^{\lceil \sqrt{n}/\sigma\rceil} \cdot Pr\left(\chi^2_{\lceil \sqrt{n}/2\sigma\rceil - 1} \le \tfrac{\sqrt{n}}{100\sigma}\right) \le 2^{\lceil \sqrt{n}/\sigma\rceil} \cdot e^{-0.7084\lceil \sqrt{n}/\sigma\rceil} \le e^{-0.015\lceil \sqrt{n}/\sigma\rceil} \le e^{-0.015\sigma^{-1}\sqrt{n}} .$$

References

P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

T.T. Cai, G. Xu, and J. Zhang. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397, 2009.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12): 4203–4215, 2005. ISSN 0018-9448.

R. Foygel and M. Drton. Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, 23:604–612, 2010.

R.H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, 23:2199–2207, 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.